3B2v8.06a/w (Dec 5 2003).51c XML:ver:5.0.1 PBR – V165 : 65004 Prod.Type: pp.33–58(col.fig.:NIL) ED: PAGN: SCAN:

Cisek, Drew & Kalaska (Eds.)
Progress in Brain Research, Vol. 165
ISSN 0079-6123

CHAPTER 4

A quantitative theory of immediate visual recognition

Thomas Serre*, Gabriel Kreiman[a], Minjoon Kouh, Charles Cadieu[b], Ulf Knoblich and Tomaso Poggio

Center for Biological and Computational Learning, McGovern Institute for Brain Research, Computer Science and Artificial Intelligence Laboratory, Brain and Cognitive Sciences Department, Massachusetts Institute of Technology, 43 Vassar Street # 46-5155B, Cambridge, MA 02139, USA

Abstract: Human and non-human primates excel at visual recognition tasks. The primate visual system exhibits a strong degree of selectivity while at the same time being robust to changes in the input image. We have developed a quantitative theory to account for the computations performed by the feedforward path in the ventral stream of the primate visual cortex. Here we review recent predictions by a model instantiating the theory about physiological observations in higher visual areas. We also show that the model can perform recognition on datasets of complex natural images at a level comparable to psychophysical measurements on human observers during rapid categorization tasks. In sum, the evidence suggests that the theory may provide a framework to explain the first 100–150 ms of visual object recognition. The model also constitutes a vivid example of how computational models can interact with experimental observations in order to advance our understanding of a complex phenomenon. We conclude by suggesting a number of open questions, predictions, and specific experiments for visual physiology and psychophysics.

Introduction

The primate visual system rapidly and effortlessly recognizes a large number of diverse objects in cluttered, natural scenes. In particular, it can easily categorize images or parts of them, for instance as an office scene or a face within that scene, and identify a specific object. This remarkable ability is evolutionarily important since it allows us to distinguish friend from foe and identify food targets in complex, crowded scenes. Despite the ease with which we see, visual recognition — one of the key issues addressed in computer vision — is quite difficult for computers. The problem of object recognition is even more difficult from the point of view of neuroscience, since it involves several levels of understanding from the information processing or computational level to circuits and biophysical mechanisms. After decades of work in different brain areas ranging from the retina to higher cortical areas, the emerging picture of how cortex performs object recognition is becoming too complex for any simple qualitative ''mental'' model.

A quantitative, computational theory can provide a much-needed framework for summarizing and integrating existing data and for planning, coordinating, and interpreting new experiments. Models are powerful tools in basic research, integrating knowledge across several levels of analysis

*Corresponding author. Tel.: +1617 253 0548; Fax:; E-mail: serre@mit.edu

[a]Current address: Children's Hospital Boston, Harvard Medical School.

[b]Current address: Redwood Center for Theoretical Neuroscience and Helen Wills Neuroscience Institute, University of California, Berkeley.

DOI: 10.1016/S0079-6123(06)65004-8

— from molecular to synaptic, cellular, systems and to complex visual behavior. In this paper, we describe a quantitative theory of object recognition in primate visual cortex that (1) bridges several levels of understanding from biophysics to physiology and behavior and (2) achieves human level performance in rapid recognition of complex natural images. The theory is restricted to the feedforward path of the ventral stream and therefore to the first 100–150 ms of visual recognition; it does not describe top-down influences, though it should be, in principle, capable of incorporating them.

In contrast to other models that address the computations in any one given brain area (such as primary visual cortex) or attempt to explain a particular phenomenon (such as contrast adaptation or a specific visual illusion), we describe here a large-scale neurobiological model that attempts to describe the basic processes across multiple brain areas. One of the initial key ideas in this and many other models of visual processing (Fukushima, 1980; Perrett and Oram, 1993; Mel, 1997; Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999) come from the pioneering physiological studies and models of Hubel and Wiesel (1962).

Following their work on striate cortex, they proposed a hierarchical model of cortical organization. They described a hierarchy of cells within the primary visual cortex: at the bottom of the hierarchy, the radially symmetric cells behave similarly to cells in the thalamus and respond best to small spots of light. Second, the simple cells which do not respond well to spots of light require bar-like (or edge-like) stimuli at a particular orientation, position, and phase (i.e., white bar on a black background or dark bar on a white background). In turn, complex cells are also selective for bars at a particular orientation but they are insensitive to both the location and the phase of the bar within their receptive fields. At the top of the hierarchy, hypercomplex cells not only respond to bars in a position and phase invariant way like complex cells, but also are selective for bars of a particular length (beyond a certain length their response starts to decrease). Hubel and Wiesel suggested that such increasingly complex and invariant object representations could be progressively built by integrating convergent inputs from lower levels. For instance, position invariance at the complex cell level could be obtained by pooling over simple cells at the same preferred orientation but at slightly different positions. The main contribution from this and other models of visual processing (Fukushima, 1980; Perrett and Oram, 1993; Mel, 1997; Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999) has been to extend the notion of hierarchy beyond V1 to extrastriate areas and show how this can explain the tuning properties of neurons in higher areas of the ventral stream of the visual cortex.

A number of biologically inspired algorithms have been described (Fukushima, 1980; LeCun et al., 1998; Ullman et al., 2002; Wersing and Koerner, 2003), i.e., systems which are only qualitatively constrained by the anatomy and physiology of the visual cortex. However, there have been very few neurobiologically plausible models (Olshausen et al., 1993; Perrett and Oram, 1993; Mel, 1997; Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999; Thorpe, 2002; Amit and Mascaro, 2003) that try to address a generic, high-level computational function such as object recognition by summarizing and integrating a large body of data from different levels of understanding. What should a general theory of biological object recognition be able to explain? It should be constrained to match data from anatomy and physiology at different stages of the ventral stream as well as human performance in complex visual tasks such as object recognition. The theory we propose may well be incorrect. Yet it represents a set of claims and ideas that deserve to be either falsified or further developed and refined.

The scope of the current theory is limited to "immediate recognition," i.e., to the first 100–150 ms of the flow of information in the ventral stream. This is behaviorally equivalent to considering "rapid categorization" tasks for which presentation times are fast and back-projections are likely to be inactive (Lamme and Roelfsema, 2000). For such tasks, presentation times do not allow sufficient time for eye movements or shifts of attention (Potter, 1975). Furthermore, EEG studies (Thorpe et al., 1996) provide evidence that the human visual system is able to solve an object detection task — determining whether a natural scene contains an animal or not — within 150 ms. Extensive evidence shows that the responses of inferior temporal (IT) cortex neurons begin 80–100 ms after onset of the visual stimulus (Perrett et al., 1992). Furthermore, the neural responses at the IT level are tuned to the stimulus essentially from response onset (Keysers et al., 2001). Recent data (Hung et al., 2005) show that the activity of small neuronal populations in IT (~100 randomly selected cells) over very short time intervals from response onset (as small as 12.5 ms) contains surprisingly accurate and robust information supporting visual object categorization and identification tasks. Finally, rapid detection tasks, e.g., animal vs. non-animal (Thorpe et al., 1996), can be carried out without top-down attention (Li et al., 2002). We emphasize that none of these rules out the use of local feedback — which is in fact used by the circuits we propose for the two main operations postulated by the theory (see section on "A quantitative framework for the ventral stream") — but suggests a hierarchical forward architecture as the core architecture underlying "immediate recognition."

We start by presenting the theory in section "A quantitative framework for the ventral stream:" we describe the architecture of a model implementing the theory, its two key operations, and its learning stages. We briefly review the evidence about the agreement of the model with single cell recordings in visual cortical areas (V1, V2, V4) and describe in more detail how the final output of the model compares to the responses in IT cortex during a decoding task that attempts to identify or categorize objects (section on "Comparison with physiological observations"). In section "Performance on natural images," we further extend the approach to natural images and show that the model performs surprisingly well in complex recognition tasks and is competitive with some of the best computer vision systems. As an ultimate and more stringent test of the theory, we show that the model predicts the level of performance of human observers on a rapid categorization task. The final section discusses the state of the theory, its limitations, a number of open questions including critical experiments, and its extension to include top-down effects and cortical back-projections.

A quantitative framework for the ventral stream

Organization of the ventral stream of visual cortex

Object recognition in cortex is thought to be mediated by the ventral visual pathway (Ungerleider and Haxby, 1994). Information from the retina is conveyed to the lateral geniculate nucleus in the thalamus and then to primary visual cortex, V1. Area V1 projects to visual areas V2 and V4, and V4 in turn projects to IT, which is the last exclusively visual area along the ventral stream (Felleman and van Essen, 1991). Based on physiological and lesion experiments in monkeys, IT has been postulated to play a central role in object recognition (Schwartz et al., 1983). It is also a major source of input to prefrontal cortex (PFC) that is involved in linking perception to memory and action (Miller, 2000).

Neurons along the ventral stream (Perrett and Oram, 1993; Logothetis and Sheinberg, 1996; Tanaka, 1996) show an increase in receptive field size as well as in the complexity of their preferred stimuli (Kobatake and Tanaka, 1994). Hubel and Wiesel (1962) first described *simple cells* in V1 with small receptive fields that respond preferentially to oriented bars. At the top of the ventral stream, IT cells are tuned to complex stimuli such as faces and other objects (Gross et al., 1972; Desimone et al., 1984; Perrett et al., 1992).

A hallmark of the cells in IT is the robustness of their firing over stimulus transformations such as scale and position changes (Perrett and Oram, 1993; Logothetis et al., 1995; Logothetis and Sheinberg, 1996; Tanaka, 1996). In addition, as other studies have shown, most neurons show specificity for a certain object view or lighting condition (Hietanen et al., 1992; Perrett and Oram, 1993; Logothetis et al., 1995; Booth and Rolls, 1998) while other neurons are view-invariant and in agreement with earlier predictions (Poggio and Edelman, 1990). Whereas view-invariant recognition requires visual experience of the specific novel object, significant position and scale invariance seems to be immediately present in the view-tuned neurons (Logothetis et al., 1995) without the need of visual experience for views *of the specific object* at different positions and scales (see also Hung et al., 2005).

In summary, the accumulated evidence points to four, mostly accepted, properties of the feedforward path of the ventral stream architecture: (a) a hierarchical build-up of invariances first to position and scale and then to viewpoint and other transformations; (b) an increasing selectivity, originating from inputs from previous layers and areas, with a parallel increase in both the size of the receptive fields and in the complexity of the optimal stimulus; (c) a basic feedforward processing of information (for "immediate recognition" tasks); and (d) plasticity and learning probably at all stages with a time scale that decreases from V1 to IT and PFC.

Architecture and model implementation

The physiological data summarized in the previous section, together with computational considerations on image invariances, lead to a theory that summarizes and extends several previously existing neurobiological models (Hubel and Wiesel, 1962; Poggio and Edelman, 1990; Perrett and Oram, 1993; Mel, 1997; Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999) and biologically motivated computer vision approaches (Fukushima, 1980; LeCun et al., 1998; Ullman et al., 2002). The theory maintains that:

One of the main functions of the ventral stream pathway is to achieve an exquisite trade-off between selectivity and invariance at the level of shape-tuned and invariant cells in IT from which many recognition tasks can be readily accomplished; the key computational issue in object recognition is to be able to finely discriminate between different objects and object classes while at the same time being tolerant to object transformations such as scaling, translation, illumination, viewpoint changes, changes in context and clutter, non-rigid transformations (such as a change of facial expression) and, for the case of categorization, also to shape variations within a class.

The underlying architecture is hierarchical, with a series of stages that gradually increase invariance to object transformations and tuning to more specific and complex *features*.

There exist at least two main functional types of units, *simple* and *complex*, which represent the result of two main operations to achieve selectivity (*S* layer) and invariance (*C* layer). The two corresponding operations are a (bell-shaped) Gaussian-like TUNING of the simple units and a MAX-like operation for invariance to position, scale, and clutter (to a certain degree) of the complex units.

Two basic operations for selectivity and invariance

The *simple S* units perform a TUNING operation over their afferents to build object-selectivity. The *S* units receive convergent inputs from retinotopically organized units tuned to *different preferred stimuli* and combine these *subunits* with a bell-shaped tuning function, thus increasing object selectivity and the complexity of the preferred stimulus. Neurons with a Gaussian-like bell-shaped tuning are prevalent across cortex. For instance, simple cells in V1 exhibit a Gaussian tuning around their preferred orientation; cells in AIT are typically tuned around a particular view of their preferred object. From the computational point of view, Gaussian-like tuning profiles may be the key in the generalization ability of the cortex. Indeed, networks that combine the activity of several units tuned with a Gaussian profile to different training examples have proved to be a powerful learning scheme (Poggio and Edelman, 1990).

The *complex C* units perform a MAX-like operation over their afferents to gain invariance to several object transformations. The complex *C* units receive convergent inputs from retinotopically organized *S* units tuned to the *same preferred stimulus* but at slightly different positions and scales and combine these subunits with a MAX-like operation, thereby introducing tolerance to scale and translation. The existence of a MAX operation in visual cortex was proposed by Riesenhuber and Poggio (1999) from theoretical arguments [and limited experimental evidence (Sato, 1989)] and was later supported experimentally in both V4 (Gawne and Martin, 2002) and V1 at the complex cell level (Lampl et al., 2004).

A gradual increase in both selectivity and invariance, to 2D transformations, as observed along the ventral stream and as obtained in the model by interleaving the two key operations, is critical for avoiding both a combinatorial explosion in the number of units and the binding problem between features. Below we shortly give idealized mathematical expressions for the operations.

Idealized mathematical descriptions of the two operations: In the following, we denote by y the response of a unit (simple or complex). The set of inputs to the cell (i.e., pre-synaptic units) are denoted with subscripts $j = 1, \ldots N$. When presented with a pattern of activity $\mathbf{x} = (x_1, \ldots, x_N)$ as input, an idealized and static description of a complex unit response y is given by:

$$y = \max_{j=1,\ldots,N} x_j \tag{1}$$

As mentioned above, for a complex cell, the inputs x_j are retinotopically organized (selected from an $m \times m$ grid of afferents with the same selectivity). For instance, in the case of a V1-like complex cell tuned to a horizontal bar, all input subunits are tuned to a horizontal bar but at slightly different positions and scales. Similarly, an idealized description of a simple unit response is given by:

$$y = exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{N}(w_j - x_j)^2\right) \tag{2}$$

σ defines the sharpness of the TUNING of the unit around its preferred stimulus corresponding to the synaptic strengths $\mathbf{w} = (w_1, \ldots, w_N)$. As for complex cells, the subunits of the simple cells are also retinotopically organized (selected from an $m \times m$ grid of possible afferents). In contrast with complex cells, the subunits of a simple cell have different selectivities to increase the complexity of the preferred stimulus. For instance, for the S_2 units, the subunits are V1-like complex cells at different preferred orientations. The response of a simple unit is maximal when the current pattern of input $\mathbf{x}$ matches exactly the synaptic weights $\mathbf{w}$ (for instance the frontal view of a face) and decreases with a bell-shaped profile as the pattern of input becomes more dissimilar (as the face is rotated away from the preferred view).

Both of these mathematical descriptions are only meant to describe the response behavior of cells at a phenomenological level. Plausible biophysical circuits for the TUNING and MAX operations have been proposed based on feedforward and/or feedback shunting inhibition combined with normalization [see Serre et al. (2005) and references therein].

Building a dictionary of shape-components from V1 to IT

The overall architecture is sketched in Fig. 1 and reflects the general organization of the visual cortex in a series of layers from V1 to IT and PFC. Colors encode the tentative correspondences between the functional primitives of the theory (right) and the structural primitives of the ventral stream in the primate visual system (Felleman and van Essen, 1991) (left, modified from Gross, 1998). Below we give a brief description of a model instantiating the theory. The reader should refer to Serre (2006) for a more complete description of the architecture and detailed parameter values.

The first stage of simple units (S_1), corresponding to the classical simple cells of Hubel and Wiesel, represents the result of the first tuning operation. Each S_1 cell is tuned in a Gaussian-like way to a bar (a gabor) of one of four possible orientations. Each of the complex units in the second layer (C_1), corresponding to the classical complex cells of Hubel and Wiesel, receives, within a neighborhood, the outputs of a group of simple units in the first layer at slightly different positions and sizes but with the same preferred orientation. The operation is a nonlinear MAX-like operation [see Eq. (1)] that increases invariance to local changes in position and scale while maintaining feature specificity.

At the next simple cell layer (S_2), the units pool the activities of several complex units (C_1) with weights dictated by the unsupervised learning stage (see below), yielding selectivity to more complex patterns such as combinations of oriented

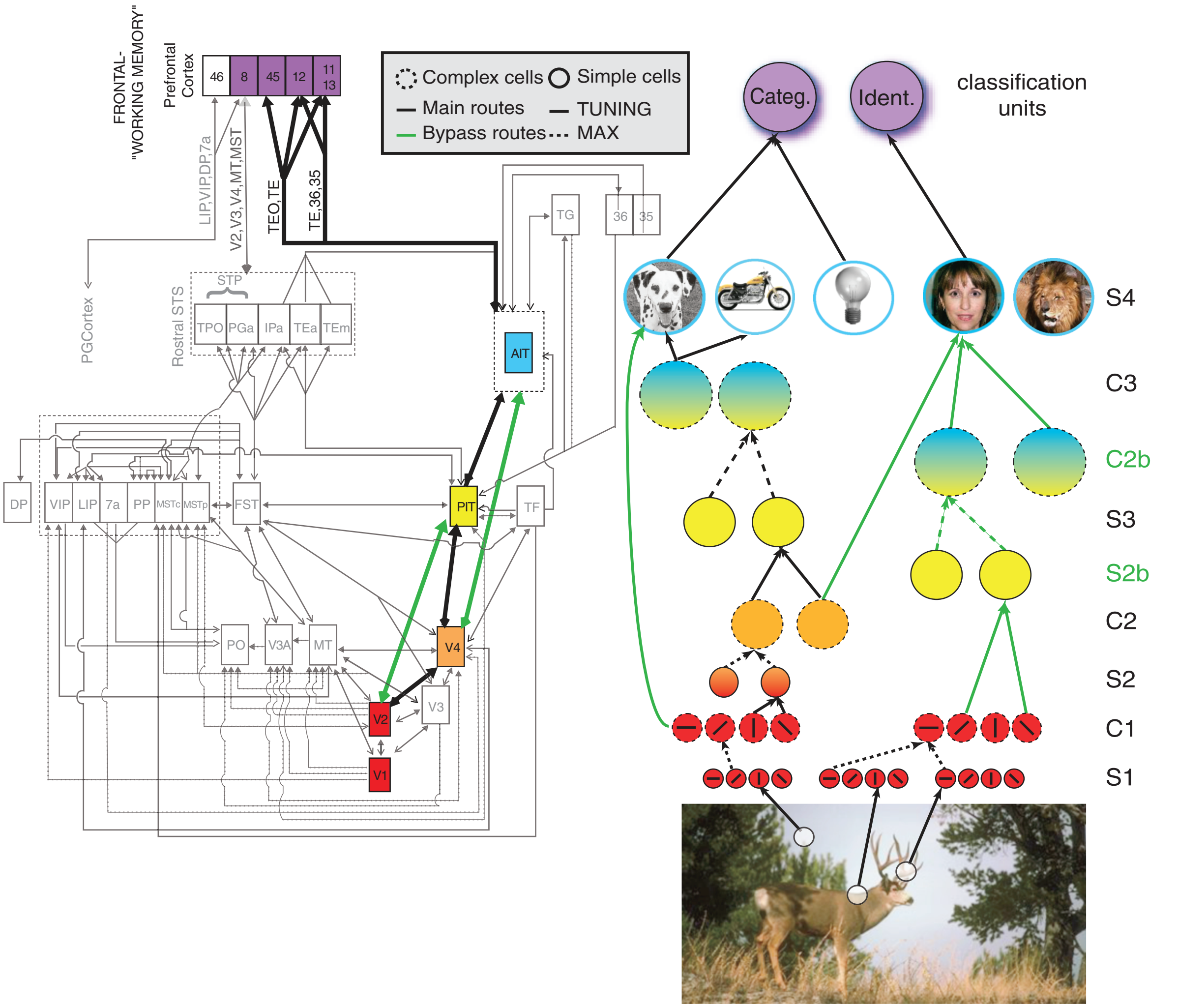

Fig. 1. Tentative mapping between structural primitives of the ventral stream in the primate visual system (Felleman and van Essen, 1991) (left) and functional primitives of the theory. The model, which is feedforward (apart from local recurrent circuits), attempts to describe the initial stage of visual processing and immediate recognition, corresponding to the output of the top of the hierarchy and to the first 150 ms in visual recognition. Colors encode the tentative correspondences between model layers and brain areas. Stages of simple cells with Gaussian-like tuning (plain circles and arrows), which provide generalization (Poggio and Bizzi, 2004), are interleaved with layers of complex units (dotted circles and arrows), which perform a MAX-like operation on their inputs and provide invariance to position and scale (pooling over scales is not shown in the figure). Both operations may be performed by the same local recurrent circuits of lateral inhibition (see text). It is important to point out that the hierarchy is probably not as strict as depicted here. In addition there may be cells with relatively complex receptive fields already in V1. The main route from the feedforward ventral pathway is denoted with black arrows while the bypass route (Nakamura et al., 1993) is denoted with green arrows. Learning in the simple unit layers from V2/V4 up to IT (including the S_4 view-tuned units) is assumed to be stimulus-driven. It only depends on task-independent visual experience-dependent tuning of the units. Supervised learning occurs at the level of the circuits in PFC (two sets of possible circuits for two of the many different recognition tasks — identification and categorization — are indicated in the figure at the level of PFC). (Adapted with permission from Serre et al., 2007a, Fig. 1.)

lines. Simple units in higher layers (S_3 and S_4) combine more and more complex features with a Gaussian tuning function [see Eq. (2)], while the complex units (C_2 and C_3) pool their afferents through a MAX-like function [see Eq. (1)], providing increasing invariance to position and scale. In the model, the two layers alternate (see Riesenhuber and Poggio, 1999). Besides the main route that follows stages along the hierarchy of the ventral stream step-by-step, there are several routes which *bypass* some of the stages, e.g., direct projections from V2 to posterior IT (bypassing V4) and from V4 to anterior IT (bypassing posterior IT cortex). In the model, such *bypass* routes correspond, for instance, to the projections from the C_1 layer to the S_{2b} and then C_{2b} layers. Altogether the various layers in the architecture — from V1 to IT — create a large and redundant dictionary of features with different degrees of selectivity and invariance.

Although the present implementation follows the hierarchy of Fig. 1, the ventral stream's hierarchy may not be as strict. For instance there may be units with relatively complex receptive fields already in V1 (Mahon and DeValois, 2001; Victor et al., 2006). A mixture of cells with various levels of selectivity has also commonly been reported in V2, V4, and IT (Tanaka, 1996; Hegdé and van Essen, 2006). In addition, it is likely that the same stimulus-driven learning mechanisms implemented for the S_2 units and above operate also at the level of the S_1 units. This may generate S_1 units with TUNING not only for oriented bars but also for more complex patterns (e.g., corners), corresponding to the combination of LGN-like, center-surround subunits in specific geometrical arrangements. Indeed it may be advantageous for circuits in later stages (e.g., task-specific circuits in PFC) to have access not only to the highly invariant and selective units of AIT but also to less invariant and simpler units such as those in V2 and V4. Fine orientation discrimination tasks, for instance, may require information from lower levels of the hierarchy such as V1. There might also be high level recognition tasks that benefit from less invariant representations.

Learning

Unsupervised developmental-like learning from V1 to IT: Various lines of evidence suggest that visual experience, both during and after development, together with genetic factors, determine the connectivity and functional properties of cells in cortex. In this work, we assume that learning plays a key role in determining the wiring and the synaptic weights for the model units. We suggest that the TUNING properties of simple units at various levels in the hierarchy correspond to learning that combinations of features appear most frequently in images. This is roughly equivalent to learning a dictionary of image patterns that appear with high probability. The wiring of the S layers depends on learning correlations of features in the image that are present at the same time (i.e., for S_1 units, the bar-like arrangements of LGN inputs, for S_2 units, more complex arrangements of bar-like subunits, etc.).

The wiring of complex cells, on the other hand, may reflect learning from visual experience to associate frequent transformations in time, such as translation and scale, of specific complex features coded by simple cells. The wiring of the C layers could reflect learning correlations *across time*: e.g., at the C_1 level, learning that afferent S_1 units with the same orientation and neighboring locations should be wired together because such a pattern often changes smoothly in time (under translation) (Földiák, 1991). Thus, learning at the S and C levels involves learning correlations present in the visual world. At present it is still unclear whether these two types of learning require different types of synaptic learning rules or not.

In the present model we have only implemented learning at the higher level S areas (beyond S_1). Connectivity at the C level was hardwired based on physiology data. The goal of this learning stage is to determine the selectivity of the S units, i.e., set the weight vector $\mathbf{w}$ (see Eq. (2)) of the units in layers S_2 and higher. More precisely, the goal is to define the basic types of units in each of the S layers, which constitute a dictionary of shape-components that reflect the statistics of natural images. This assumption follows the notion that the visual system, through visual experience and evolution, may be adapted to the statistics of its natural environment (Barlow, 1961). Details about the learning rule can be found in (Serre, 2006).

Supervised learning of the task-specific circuits from IT to PFC: For a given task, we assume that a particular program or routine is set up somewhere beyond IT (possibly in PFC (Freedman et al., 2002; Hung et al., 2005), but the exact locus may depend on the task). In a passive state (no specific visual task is set) there may be a default routine running (perhaps the routine: what is out there?). Here we think of a particular classification routine as a particular PFC-like unit that combines the activity of a few hundred S_4 units tuned to produce a high response to examples of the target object and low responses to distractors. While learning in the S layers is stimulus-driven, the PFC-like classification units are trained in a supervised way. The concept of a classifier that takes its inputs from a few broadly tuned example-based units is a learning scheme that is closely related to Radial Basis Function (RBF) networks (Poggio and Edelman, 1990), which are among the most powerful classifiers in terms of generalization ability. Computer simulations have shown the plausibility of this scheme for visual recognition and its quantitative consistency with many data from physiology and psychophysics (Poggio and Bizzi, 2004).

In the model, the response of a PFC-like *classification* unit with input weights $\mathbf{c} = (c_1, \ldots, c_n)$ is given by:

$$f(\mathbf{x}) = \sum_i c_i K(\mathbf{x^i}, \mathbf{x})$$

$$\text{where} \quad K(\mathbf{x^i}, \mathbf{x}) = exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j^i - x_j)^2\right) \qquad (3)$$

$K(\mathbf{x^i},\mathbf{x})$ characterizes the activity of the i^{th} S_4 unit, tuned to the training example $\mathbf{x^i}$, in response to the input image $\mathbf{x}$ and was obtained by replacing the weight vector $\mathbf{w}$ in Eq. (2) by the training example $\mathbf{x^i}$ (i.e., $\mathbf{w} = \mathbf{x^i}$). The superscript i indicates the index of the image in the training set and the subscript j indicates the index of the pre-synaptic unit. Supervised learning at this stage involves adjusting the synaptic weights $\mathbf{c}$ to minimize the overall classification error on the training set (see Serre, 2006).

Comparison with physiological observations

The quantitative implementation of the model, as described in the previous section, allows for direct comparisons between the responses of units in the model and electrophysiological recordings from neurons in the visual cortex. Here we illustrate this approach by directly comparing the model against recordings from the macaque monkey area V4 and IT cortex while the animal was passively viewing complex images.

Comparison of model units with physiological recordings in the ventral visual cortex

The model includes several layers that are meant to mimic visual areas V1, V2, V4, and IT cortex (Fig. 1). We directly compared the responses of the model units against electrophysiological recordings obtained throughout all these visual areas. The model is able to account for many physiological observations in early visual areas. For instance, at the level of V1, model units agree with the tuning properties of cortical cells in terms of frequency and orientation bandwidth, as well as peak frequency selectivity and receptive field sizes (see Serre and Riesenhuber, 2004). Also in V1, we observe that model units in the C_1 layer can explain responses of a subpopulation of complex cells obtained upon presenting two oriented bars within the receptive field (Lampl et al., 2004). At the level of V4, model C_2 units exhibit tuning for complex gratings (based on the recordings from Gallant et al., 1996), and curvature (based on Pasupathy and Connor, 2001), as well as interactions of multiple dots (based on Freiwald et al., 2005) or the simultaneous presentation of two-bar stimuli [based on Reynolds et al. (1999), see Serre et al. (2005) for details].

Here we focus on one comparison between C_2 units and the responses of V4 cells. Figure 2 shows the side-by-side comparison between a model C_2 unit and V4 cell responses to the presentation of one-bar and two-bar stimuli. As in (Reynolds et al., 1999) model units were presented with either (1) a *reference* stimulus alone (an oriented bar at position 1, see Fig. 2A), (2) a *probe* stimulus alone

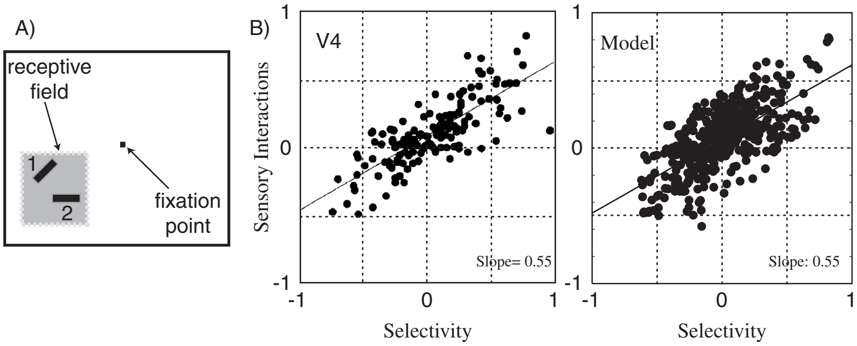

Fig. 2. A quantitative comparison between model C_2 units and V4 cells. (A) Stimulus configuration (adapted with permission from Reynolds et al., 1999, Fig. 1A): The stimulus in position 1 is denoted as the reference and the stimulus in position 2 as the probe. As in Reynolds et al. (1999) we computed a *selectivity* index (which indicates how selective a cell is to an isolated stimulus in position 1 vs. position 2 alone) and a *sensory interaction* index (which indicates how selective the cell is to the paired stimuli vs. the reference stimulus alone) (see text and Serre et al., 2005 for details). (B) Side-by-side comparison between V4 neurons (left, adapted with permission from Reynolds et al., 1999, Fig. 5) while the monkey attends away from the receptive field location and C_2 units (right). Consistent with the physiology, the addition of a second stimulus in the receptive field of the C_2 unit moves the response of the unit toward that of the second stimulus alone, i.e., the response to the clutter condition lies between the responses to the individual stimuli.

(an oriented bar at position 2), or (3) both a reference and a probe stimulus simultaneously. We used stimuli of 16 different orientations for a total of $289 = (16+1)^2$ total stimulus combinations for each unit [see Serre et al. (2005) for details]. Each unit's response was normalized by the maximal response of the unit across all conditions. As in Reynolds et al. (1999) we computed a *selectivity* index as the normalized response of the unit to the reference stimulus minus the normalized response of the unit to one of the probe stimuli. This index was computed for each of the probe stimuli, yielding 16 selectivity values for each model unit. This selectivity index ranges from -1 to $+1$, with negative values indicating that the reference stimulus elicited the stronger response, a value of 0 indicating identical responses to reference and probe, and positive values indicating that the probe stimulus elicited the strongest response. We also computed a *sensory interaction* index that corresponds to the normalized response to a pair of stimuli (the reference and a probe) minus the normalized response to the reference alone. The selectivity index also takes on values from -1 to $+1$. Negative values indicate that the response to the pair is smaller than the response to the reference stimulus alone (i.e., adding the probe stimulus suppresses the neuronal response). A value of 0 indicates that adding the probe stimulus has no effect on the neuron's response while positive values indicate that adding the probe increases the neuron's response.

As shown in Fig. 2B, model C_2 units and V4 cells behave very similarly to the presentation of two stimuli within their receptive field. Indeed the slope of the *selectivity* vs. *sensory interaction* indices is ~0.5 for both model units and cortical cells. That is, at the population level, presenting a preferred and a non-preferred stimulus together produces a neural response that falls between the neural responses to the two stimuli individually, sometimes close to an average.[1] We have found that such a "clutter effect" also happens higher up in the hierarchy at the level of IT (see Serre et al., 2005). Since normal vision operates with many objects appearing within the same receptive fields and embedded in complex textures (unlike the artificial experimental setups), understanding the behavior of neurons under clutter conditions is important and warrants more experiments (see

[1] We only compare the response of the model units to V4 neurons when the monkey is attending away from the receptive field location of the neuron. When the animal attends at the location of the receptive field the response to the pairs is shifted towards the response to the attended stimulus.

later section ‘‘Performance on natural images’’ and section ‘‘A quantitative framework for the ventral stream’’).

In sum, the model can capture many aspects of the physiological responses of neurons along the ventral visual stream from V1 to IT cortex (see also Serre et al., 2005).

Decoding object information from IT and model units

We recently used a simple linear statistical classifier to quantitatively show that we could accurately, rapidly, and robustly decode visual information about objects from the activity of small populations of neurons in anterior IT cortex (Hung et al., 2005). In collaboration with Chou Hung and James DiCarlo at MIT, we observed that a binary response from the neurons (using small bins of 12.5 ms to count spikes) was sufficient to encode information with high accuracy. This robust visual information, as measured by our classifiers, could in principle be decoded by the targets of IT cortex such as PFC to determine the class or identity of an object (Miller, 2000). Importantly, the population response generalized across object positions and scales. This scale and position invariance was evident even for novel objects that the animal never observed before (see also Logothetis et al., 1995). The observation that scale and position invariance occurs for novel objects strongly suggests that these two forms of invariance do not require multiple examples of each specific object. This should be contrasted with other forms of invariance, such as robustness to depth rotation, which requires multiple views in order to be able to generalize (Poggio and Edelman, 1990).

Read-out from C_{2b} units is similar to decoding from IT neurons

We examined the responses of the model units to the same set of 77 complex object images seen by the monkey. These objects were divided into eight possible categories. The model unit responses were divided into a training set and a test set. We used a one-versus-all approach, training eight binary classifiers, one for each category against the rest of the categories, and then taking the classifier prediction to be the maximum among the eight classifiers (for further details, see Hung et al., 2005; Serre et al., 2005). Similar observations were made when trying to identify each individual object by training 77 binary classifiers. For comparison, we also tried decoding object category from a random selection of model units from other layers of the model (see Fig. 1). The input to the classifier consisted of the responses of randomly selected model units and the labels of the object categories (or object identities for the identification task). Data from multiple units were concatenated assuming independence.

We observed that we could accurately read out the object category and identity from model units. In Fig. 3A, we compare the classification performance, for the categorization task described above, between the IT neurons and the C_{2b} model units. In agreement with the experimental data from IT, units from the C_{2b} stage of the model yielded a high level of performance ($>70\%$ for 100 units; where chance was 12.5%). We observed that the physiological observations were in agreement with the predictions made by the highest layers in the model (C_{2b}, S_4) but not by earlier stages (S_1 through S_2). As expected, the layers from S_1 through S_2 showed a weaker degree of scale and position invariance.

The classification performance of S_{2b} units (the input to C_{2b} units, see Fig. 1) was qualitatively close to the performance of local field potentials (LFPs) in IT cortex (Kreiman et al., 2006). The main components of LFPs are dendritic potentials and therefore LFPs are generally considered to represent the dendritic input and local processing within a cortical area (Mitzdorf, 1985; Logothetis et al., 2001). Thus, it is tempting to speculate that the S_{2b} responses in the model capture the type of information conveyed by LFPs in IT. However, care should be taken in this interpretation as the LFPs constitute an aggregate measure of the activity over many different types of neurons and large areas. Further investigation of the nature of the LFPs and their relation with the spiking

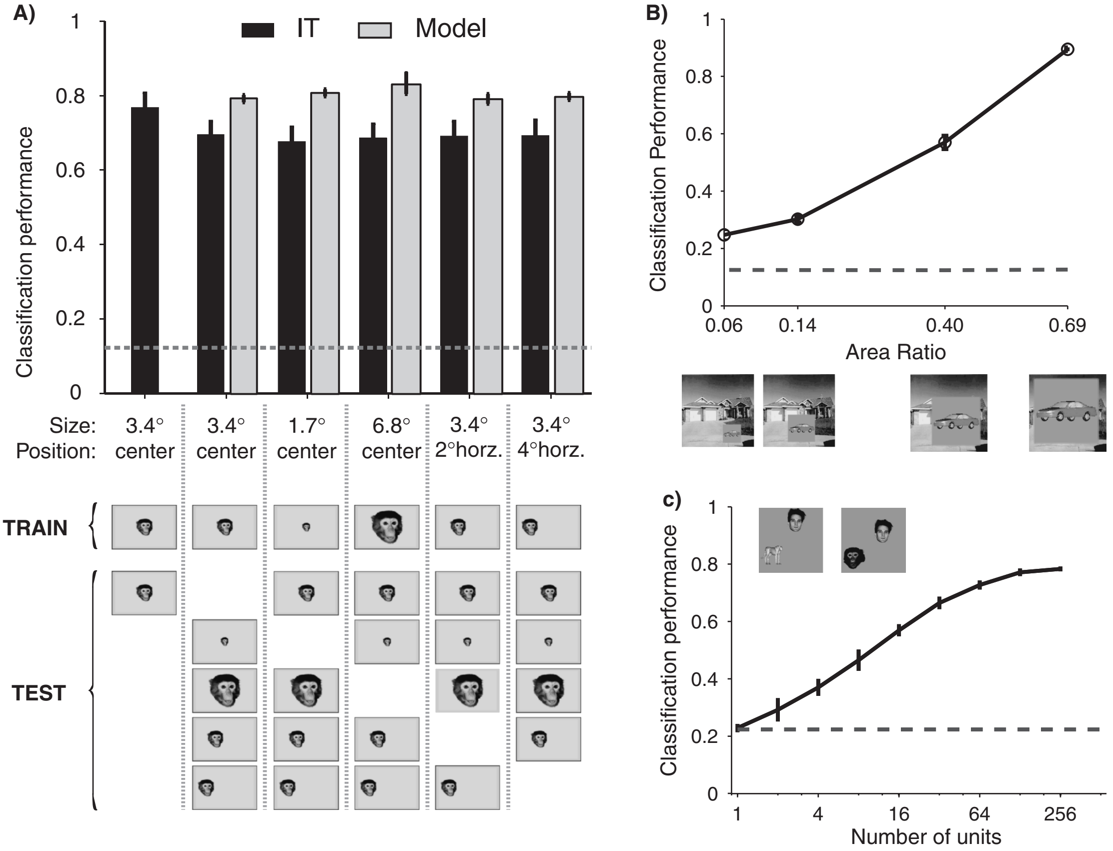

Fig. 3. (A) Classification performance based on the spiking activity from IT neurons (black) and C_{2b} units from the model (gray). The performance shown here is based on the categorization task where the classifier was trained based on the category of the object. A linear classifier was trained using the responses to the 77 objects at a single scale and position (shown for one object by "TRAIN"). The classifier performance was evaluated using shifted or scaled versions of the same 77 objects (shown for one object by "TEST"). During training, the classifier was never presented with the unit responses to the shifted or scaled objects. The left-most column shows the performance for training and testing on separate repetitions of the objects at the same standard position and scale (this is shown only for the IT neurons because there is no variability in the model which is deterministic). The second bar shows the performance after training on the standard position and scale (3.4°, center of gaze) and testing on the shifted and scaled images. The dashed horizontal line indicates chance performance (12.5%, one out of eight possible categories). Error bars show standard deviations over 20 random choices of the units used for training/testing. (B) Classification performance for reading out object category as a function of the relative size (area ratio) of object to background. Here the classifier was trained using the responses of 256 units to the objects presented in cluttered backgrounds. The classifier performance was evaluated using the same objects embedded in different backgrounds. The horizontal dashed line indicates chance performance obtained by randomly shuffling the object labels during training. (C) Classification performance for reading out object category in the presence of two objects. We exhaustively studied all possible pairs using the same 77 objects as in part A (see two examples on the upper left part of the figure). The classifier was trained with images containing two objects and the label corresponded to the category of one of them. During testing, the classifier's prediction was considered to be a hit if it correctly categorized either of the objects present in the image. The dashed line indicates change performance obtained by randomly assigning object labels during training.

responses could help unravel the transformations that take place across cortical layers.

The pattern of errors made by the classifier indicates that some groups were easier to discriminate than others. This was also evident in the correlation matrix of the population responses between all pairs of pictures (Hung et al., 2005; Serre et al., 2005). The units yielded similar responses to stimuli that looked alike at the pixel level. The performance of the classifier for categorization dropped significantly upon arbitrarily defining the categories as random groups of pictures.

We also tested the ability of the model to generalize to novel stimuli not included in the training set. The performance values shown in Fig. 3A are based on the responses of model units to single stimulus presentations that were not included in the classifier training and correspond to the results obtained using a linear classifier. Although the way in which the weights were learned (using a support vector machine classifier) is probably very different in biology (see Serre, 2006); once the weights are established the linear classification boundary could very easily be implemented by neuronal hardware [see Eq. (3)]. Therefore, the recognition performance provides a lower bound to what a real downstream unit (e.g., in PFC) could, in theory, perform on a single trial given input consisting of a few spikes from the neurons in IT cortex. Overall, we observed that the population of C_{2b} model units yields a read-out performance level that is very similar to the one observed from a population of IT neurons.

Extrapolation to larger object sets

One of the remarkable aspects of primate visual recognition is the large number of different objects that can be identified. Although the exact limits are difficult to estimate, coarse estimates suggest that it is possible to visually recognize on the order of 10^4 different concepts (Biederman, 1987). The physiological recordings were necessarily limited to a small set of objects due to time constraints during a recording session. Here we show that this type of encoding can extrapolate to reading out object category in a set consisting of 787 objects divided into 20 categories (the physiological observations and the model results discussed above were based on 77 objects divided into 8 categories).

The population of C_{2b} units conveyed information that could be decoded to indicate an object's category across novel objects. The classifier was trained with objects from 20 possible categories presented at different random locations and the test set included novel objects never seen before by the classifier but belonging to the same categories. These results show that a relatively small neuronal population can in principle support object recognition over large object sets. Similar results were obtained in analogous computer vision experiments using an even larger set known as the *Caltech-101* object dataset (Serre et al., 2007b) where the model could perform object categorization among 101 categories. Other investigators have also used models that can extrapolate to large numbers of objects (Valiant, 2005) or suggested that neuronal populations in IT cortex can also extrapolate to many objects (Abbott et al., 1996; Hung et al., 2005).

The number of objects (or classes) that can be decoded at a given level of accuracy grows approximately as an exponential function of the number of units. Even allowing for a strong redundancy in the number of units coding each type of feature, these results suggest that networks of thousands of units could display a very large capacity. Of course the argument above relies on several assumptions that could well be wrong. However, at the very least, these observations suggest that there do not seem to be any obvious capacity limitations for hierarchical models to encode realistically large numbers of objects and categories.

Robustness in object recognition

Many biological sources of noise could affect the encoding of information. Among the most drastic sources of noise are synaptic failures and neuronal death. To model this, we considered the performance of the classifier after randomly deleting a substantial fraction of the units during testing. As shown for the experimental data in Hung et al. (2005), the classifier performance was very robust to this source of noise.

As discussed in the introduction, one of the main achievements of visual cortex is the balance of invariance *and* selectivity. Two particularly important forms of invariance are the robustness to changes in scale and position of the images. In order to analyze the degree of invariance to scale and position changes, we studied the responses of units at different stages of the model to scaled (0.5 × and 2 ×) and translated (2° and 4°) versions of the images. The earlier stages of the model show a poor read-out performance under these transformations, but the performance of the C_{2b} stage is quite robust to these transformations as shown in Fig. 3A, in good agreement with the experimental data (Hung et al., 2005).

We also observed that the population response could extrapolate to novel objects within the same categories by training the classifier on the responses to 70% of the objects and testing its performance on the remaining 30% of the objects (Serre et al., 2005). This suggests another dimension of robustness, namely, the possibility of learning about a category from some exemplars and then extrapolating for novel objects within the same category.

The results shown above correspond to randomly selecting a given number of units to train and test the classifier. The brain could be wired in a very specific manner so that only the neurons highly specialized for a given task project to the neurons involved in decoding the information for that task. Preselecting the units (e.g., using those yielding the highest signal-to-noise ratio) yields similar results while using a significantly smaller number of units. Using a very specific set of neurons (instead of randomly pooling from the population and using more neurons for decoding) may show less robustness to neuronal death and spike failures. The bias toward using only a specific subset of neurons could be implemented through selection mechanisms including attention. For example, when searching for the car keys, the weights from some neurons could be adjusted so as to increase the signal-to-noise ratio for specific tasks. This may suggest that other concomitant recognition tasks would show weaker performance. In this case, the selection mechanisms take place before recognition by biasing specific populations for certain tasks.

Recognition in clutter

The decoding experiments described above as well as a large fraction of the studies reported in the literature, involve the use of well-delimited single objects on a uniform background. This is quite remote from natural vision where we typically encounter multiple objects embedded in different backgrounds, with potential occlusions, changes in illumination, etc.

Ultimately, we would like to be able to read out information from IT or from model units under natural vision scenarios in which an everyday life image can be presented and we can extract from the population activity the same type and quality of information that a human observer can (in a flash). Here we show the degree of decoding robustness of objects that are embedded in complex backgrounds (see also section "Performance on natural images" describing the performance of the model in an animal vs. non-animal categorization task using objects embedded in complex backgrounds).

We presented the same 77 objects used in Fig. 3A overlayed on top of images containing complex background scenes (Fig. 3B). We did not attempt to make the resulting images realistic or meaningful in any way. While cognitive influences, memory, and expectations play a role in object recognition, these high-level effects are likely to be mediated by feedback biasing mechanisms that would indicate that a monitor is more likely to be found on an office desk than in the jungle. However, the model described here is purely feedforward and does not include any of these potential biasing mechanisms. We used four different relative sizes of object-to-background (ratio of object area to whole image area) ranging from 6% to 69%. The latter condition is very similar to the single object situation analyzed above, both perceptually and in terms of the performance of the classifier. The smaller relative size makes it difficult to detect the object at least in some cases when it is not salient (see also section "Performance on natural images").

The classifier was trained on all objects using 20% of the background scenes and performance was evaluated using the same objects presented on the remaining novel background scenes (we used a total of 98 complex background scenes with photographs of outdoor scenes). The population of C_{2b} units allowed us to perform both object recognition (Fig. 3B) and identification significantly above chance in spite of the background. Performance depended quite strongly on the relative image size (Fig. 3B). The largest size (69%) yielded results that were very close to the single isolated object results discussed above (cf. Fig. 3A). The small relative image size (6%) yielded comparatively lower results but the performance of C_{2b} units was still significantly above chance levels both for categorization and identification.

Recognizing (and searching for) small objects embedded in a large complex scene (e.g., searching for the keys in your house), constitutes an example of a task that may require additional resources. These additional resources may involve serial attention that is likely to be dependent on feedback connections. Therefore, the model may suggest tasks and behaviors that require processes that are not predominantly feedforward.

Reading-out from images containing multiple objects

In order to further explore the mechanisms for representing information about an object's identity and category in natural scenes, we studied the ability to read out information from the model units upon presentation of more than one object. We presented two objects simultaneously in each image (Fig. 3C). During testing, the classifier was presented with images containing multiple objects. We asked two types of questions: (1) what is the most likely object in the image? and (2) what are all the objects present in the image?

Training was initially performed with single objects. Interestingly, we could also train the classifier using images containing multiple objects. In this case, for each image, the label was the identity (or category) of one of the objects (randomly chosen so that the overall training set had the same number of examples for each of the objects or object categories). This is arguably a more natural situation in which we learn about objects since we rarely see isolated objects. However, it is possible that attentional biases to some extent ‘‘isolate’’ an object (e.g., when learning about an object with an instructor that points to it).

In order to determine the most likely object present in the image (question 1, above), the classifier's prediction was considered to be a hit if it correctly predicted either one of the two objects presented during testing. The population of C_{2b} model units yielded very high performance reaching more than 90% both for categorization and identification with the single object training and reaching more than 80% with the multiple object training. Given that in each trial there are basically two possibilities to get a hit, the chance levels are higher than the ones reported in Fig. 3A. However, it is clear that the performance of the C_{2b} population response is significantly above chance indicating that accurate object information can be read-out even in the presence of another object. We also extended these observations to 3 objects and to 10 objects (Serre et al., 2005), obtaining qualitatively similar conclusions.

Ultimately, we would like to be able to understand an image in its entirety, including a description of all of its objects. Therefore, we asked a more difficult question by requiring the classifier to correctly predict all the objects (or all the object categories) present in the image. During perception, human observers generally assume that they can recognize and describe every object in an image during a glimpse. However, multiple psychophysics studies suggest that this is probably wrong. Perhaps one of the most striking demonstrations of this fallacy is the fact that sometimes we can be oblivious to large changes in the images (see Simons and Rensink, 2005). What is the capacity of the representation at-a-glance? There is no consensus answer to this question but some psychophysical studies suggest that only a handful of objects can be described in a brief glimpse of an image (on the order of five objects). After this first glance, eye movements and/or attentional shifts may be required to further describe an image. We continue here referring to this rapid vision scenario and we strive to explain our perceptual capabilities during the glance using the model. Thus, the goal is to be able to fully describe a set of about five objects that can be simultaneously presented in multiple backgrounds in a natural scenario.

For this purpose, we addressed our second question by taking the two most likely objects (or object categories) given by the two best classifier predictions (here the number of objects was hard-wired). A hit from the classifier output was defined as a perfect match between these predictions and the two objects present in the image. This task is much more difficult (compared to the task where the goal is to categorize or identify *any* of the objects in the image). The performance of the classifier was also much smaller than the one reported for the single-object predictions. However, performance was significantly above chance, reaching almost 40% for categorization (chance $= 0.0357$) and almost 8% for identification (chance $= 3.4 \times 10^{-4}$).

Similar results were obtained upon reading out the category or identity of all objects present in the image in the case of 3-object and 10-object images. Briefly, even in images containing 10 objects, it is possible to reliably identify one arbitrary object significantly above chance from the model units. However, the model performance in trying to describe all objects in the image drops drastically with multiple objects to very low levels for 4–5 objects.

In summary, these observations suggest that it is possible to recognize objects from the activity of small populations of IT-like model units under natural situations involving complex backgrounds and several objects. The observations also suggest that, in order to fully describe an image containing many objects, eye movements, feedback, or other additional mechanisms may be required.

Performance on natural images

For a theory of visual cortex to be successful, it should not only mimic the response properties of neurons and the behavioral response of the system to artificial stimuli like the ones typically used in physiology and psychophysics, but should also be able to perform complex categorization tasks in a real-world setting.

Comparison between the model and computer vision systems

We extensively tested the model on standard computer vision databases for comparison with several state-of-the-art AI systems (see Serre, 2006; Serre et al., 2007b, for details). Such real-world image datasets tend to be much more challenging than the typical ones used in a neuroscience lab. They usually involve different object categories and the systems that are evaluated have to cope with large variations in shape, contrast, clutter, pose, illumination, size, etc. Given the many specific biological constraints that the theory had to satisfy (e.g., using only biophysically plausible operations, receptive field sizes, range of invariances, etc.), it was not clear how well the model implementation described in section "A quantitative framework for the ventral stream" would perform in comparison to systems that have been heuristically engineered for these complex tasks.

Surprisingly we found that the model is capable of recognizing complex images (see Serre et al., 2007b). For instance, the model performs at a level comparable to some of the best existing systems on the *CalTech-101* image database of 101 object categories (Fei-Fei et al., 2004) with a recognition rate of ~55% [chance level <1%, see Serre et al. (2007b) and also the extension by Mutch and Lowe (2006)].[2] Additionally, Bileschi and Wolf have developed an automated real-world Street Scene recognition system (Serre et al., 2007b) based in part on the model described in section "A quantitative framework for the ventral stream." The system is able to recognize seven different object categories (cars, bikes, pedestrians, skies, roads, buildings, and trees) from natural images of street scenes despite very large variations in shape (e.g., trees in summer and winter, SUVs as well as compact cars under any view point).

[2] These benchmark evaluations relied on an earlier partial implementation of the model which only included the bypass route from $S_1 \rightarrow C_{2b}$.

Comparison between the model and human observers

Finally, we tested whether the level of performance achieved by the model was sufficient to account for the level of performance of human observers. To test this hypothesis, in the same way as an experimental test of Newton's second law requires choosing a situation in which friction is negligible, we looked for an experimental paradigm in which recognition has to be fast and cortical back-projections are likely to be inactive. Ultra-rapid object categorization (Thorpe et al., 1996) likely depends only on feedforward processing (Thorpe et al., 1996; Keysers et al., 2001; Thorpe and Fabre-Thorpe, 2001; Li et al., 2002; VanRullen and Koch, 2003) and thus satisfies our criterion. Here we used a backward masking paradigm (Bacon-Mace et al., 2005) in addition to the rapid stimulus presentation to try to efficiently block recurrent processing and cortical feedback loops (Enns and Di Lollo, 2000; Lamme and Roelfsema, 2000; Breitmeyer and Ogmen, 2006).

Human observers can discriminate a scene that contains a particular prominent object, such as an animal or a vehicle, after only 20 ms of exposure. Evoked response potential components related to either low-level features of the image categories (e.g., animal or vehicles) or to the image status (animal present or absent) are available at 80 and 150 ms respectively. These experimental results establish a lower bound on the latency of visual categorization decisions made by the human visual system, and suggest that categorical decisions can be implemented within a feedforward mechanism of information processing (Thorpe et al., 1996; Keysers et al., 2001; Thorpe and Fabre-Thorpe, 2001; Li et al., 2002; VanRullen and Koch, 2003).

Predicting human performance during a rapid categorization task

In collaboration with Aude Oliva at MIT, we tested human observers on a rapid animal vs. non-animal categorization task [see Serre et al. (2007a), for details]. The choice of the animal category was motivated by the fact that (1) it was used in the original paradigm by Thorpe et al. (1996) and (2) animal photos constitute a rich class of stimuli exhibiting large variations in texture, shape, size, etc. providing a difficult test for a computer vision system.

We used an image dataset that was collected by Antonio Torralba and Aude Oliva and consisted of a balanced set of 600 animal and 600 non-animal images (see Torralba and Oliva, 2003). The 600 animal images were selected from a commercially available database (Corel Photodisc) and grouped into four categories, each category corresponding to a different viewing-distance from the camera: *heads* (close-ups), *close-body* (animal body occupying the whole image), *medium-body* (animal in scene context), and *far-body* (small animal or groups of animals in larger context). One example from each group is shown in Fig. 4.

To make the task harder and prevent subjects from relying on low-level cues such as image-depth, the 600 distractor images were carefully selected to match each of the four viewing-distances. Distractor images were of two types (300 of each): artificial or natural scenes [see Serre et al. (2007a), for details].

During the experiment, images were briefly flashed for 20 ms, followed by an inter-stimulus interval (i.e., a blank screen) of 30 ms, followed by a mask (80 ms, 1/f noise). This is usually considered a long stimulus onset asynchrony ($SOA = 50$ ms) for which human observers are close to ceiling performance (Bacon-Mace et al., 2005). On the other hand, based on latencies in visual cortex, such an *SOA* should minimize the possibility of feedback and top-down effects in the task: we estimated from physiological data (see Serre et al., 2007a) that feedback signals from say, V4 to V1 or IT/PFC to V4, should not occur earlier than 40–60 ms after stimulus onset. Human observers ($n_h = 24$) were asked to respond as fast as they could to the presence or absence of an animal in the image by pressing either of the two keys.

Before we could evaluate the performance of the model, the task-specific circuits from IT to PFC (see section on ''A quantitative framework for the ventral stream'') had to be trained. These task-specific circuits correspond to a simple linear

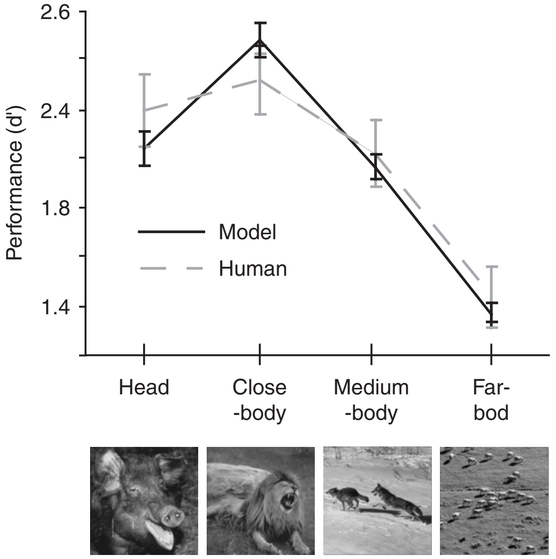

Fig. 4. Comparison between the model and human observers. Images showed either an animal embedded in a natural background or a natural scene without any animals. Images were flashed for 20 ms followed by a 30 ms blank and a 80 ms mask. Human observers or the model were queried to respond indicating whether an animal was present or not. The figure shows the accuracy as d' (the higher the value of the d', the better the performance), for the model (red) and humans (blue) across 1200 animal and non-animal stimuli. The model is able to predict the level of performance of human observers (overall 82% for the model vs. 80% for human observers). For both the model and human observers the level of performance is highest on the close-body condition and drops gradually as the amount of clutter increases in the image from close-body to medium-body and far-body. (Adapted with permission from Serre et al., 2007a, Fig. 3A.)

classifier that reads out the activity of a population of high level model units analogous to recordings from anterior IT cortex (see section on ''Comparison with physiological observations''). The training for these task-specific circuits was done by using ($n_m = 20$) random splits of the 1200 stimuli into a training set of 600 images and a test set of 600 images. For each split, we learned the synaptic weights of the task-specific circuits of the model by minimizing the error on the training set (see Serre et al., 2007a) and evaluated the model performance on the test set. The reported performance corresponds to the average performance from the random runs.

The performance of the model and of human observers was very similar (see Fig. 4). As for the model, human observers performed best on ''close-body'' views and worst on ''far-body'' views. An intermediate level of performance was obtained for ''head'' and ''medium-far'' views. Overall no significant difference was found between the level of performance of the model and human subjects. Interestingly, the observed dependency between the level of performance and the amount of clutter in the images (which increases from the close-body to the far-body condition) for both human observers and the model seems consistent with the read-out experiment from IT neurons (for both the model and human observers) as described in section ''Comparison with physiological observations.''

Importantly, lower stages of the model (C_1 units) alone could not account for the results (see Serre et al., 2007a). Additionally, performing the equivalent of the lesioning of V4 in the model (i.e., leaving the bypass routes (C_{2b} units as the only source of inputs to the final classifier), see Fig. 1), also resulted in a significant loss in performance (this was true even after retraining the task-specific circuits thus accounting for a ''recovery'' period). This lesion experiment suggests that the large dictionary of shape-tuned units in the model (from V1 to IT) with different levels of complexity and invariance learned from natural images is the key in explaining the level of performance.

Beyond comparing levels of performance, we also performed an image-by-image comparison between the model and human observers. For this comparison, we defined an index of ''animalness'' for each individual image. For the model, this index was computed by calculating the percentage of times each image was classified as an animal (irrespective of its true label) for each random run ($n_m = 20$) during which it was presented as a test image. For human observers we computed the number of times each individual image was classified as an animal by each observer ($n_h = 24$). This index measures the confidence of either the model ($n_m = 20$) or human observers ($n_h = 24$) in the presence of an animal in the image. A percentage of 100% (correspondingly 0%) indicates a very high level of confidence in the presence (absence) of an animal. The level of correlation for the animalness index between the model and human observers was 0.71, 0.84, 0.71, and 0.60 for heads, close-body, medium-body, and far-body respectively ($p < 0.01$ for testing the hypothesis of no correlation against the alternative that there is a non-zero correlation). This suggests that the model and human observers tend to produce consistent responses on individual images.

Additionally, to further challenge the model, we looked at the effect of image orientation (90° and 180° in-the-plane rotation): Rousselet et al. (2003) previously suggested that the level of performance of human observers during a rapid categorization task tends to be robust to image rotation. We found that the model and human observers exhibited a similar degree of robustness (see Serre et al., 2007a). Interestingly, the good performance of the model on rotated images was obtained without the need for retraining the model. This suggests that according to the dictionary of shape-tuned units from V1 to IT in the model (and presumably in visual cortex), an image of a rotated animal is more similar to an image of an upright animal than to distractors. In other words, a small image patch of a rotated animal is more similar to a patch of an upright animal than to a patch of image from a distractor.

Discussion: feedforward vs. feedback processing

As discussed earlier, an important assumption for the experiment described above is that with an *SOA* 50 ms, the mask leaves sufficient time to process the signal and estimate firing rates at each stage of the hierarchy (i.e., 20–50 ms, see Tovee et al., 1993; Rolls et al., 1999; Keysers et al., 2001; Thorpe and Fabre-Thorpe, 2001; Hung et al., 2005), yet selectively blocks top-down signals [e.g., from IT or PFC to V4 that we estimated to be around 40–60 ms, see Serre et al. (2007a) for a complete discussion]. The prediction is thus that the feedforward system should: (1) outperform human observers for very short *SOAs* (i.e., under 50 ms when there is not enough time to reliably perform local computations or estimate firing rates within visual areas), (2) mimic the level of performance of human observers for *SOAs* around 50 ms such that there is enough time to reliably estimate firing rates within visual areas but not enough time for back-projections from top-down to become active, and (3) underperform human observers for long *SOAs* (beyond 60 ms) such that feedbacks are active.

We thus tested the influence of the mask onset time on visual processing with four experimental conditions, i.e., when the mask followed the target image (a) without any delay (with an *SOA* of 20 ms), (b) with an *SOA* of 50 ms (corresponding to an inter-stimulus interval of 30 ms), (c) with an *SOAs* of 80 ms, or (d) never ("no-mask" condition). For all four conditions, the target presentation was fixed to 20 ms as before. As expected, the delay between the stimulus and the mask onset modulates the level of performance of the observers, improving gradually from the 20 ms *SOA* condition to the no-mask condition. The performance of the model was superior to the performance of human observers for the *SOA* of 20 ms. The model closely mimicked the level of performance of human observers for the 50 ms condition (see Fig. 4). The implication would be that, under these conditions, the present feedforward version of the model already provides a satisfactory description of information processing in the ventral stream of visual cortex. Human observers however outperformed the model for the 80 ms *SOA* and the no-mask condition.

Discussion

General remarks about the theory

We have developed a quantitative model of the feedforward pathway of the ventral stream in visual cortex — from cortical area V1 to V2 to V4 to IT and PFC — that captures its ability to learn visual tasks, such as identification and categorization of objects from images. The quantitative nature of the model has allowed us to directly compare its performance against experimental observations at different scales and also against current computer vision algorithms. In this paper we have focused our discussion on how the model can explain experimental results from visual object recognition within short times at two very different levels of analysis: human psychophysics and physiological recordings in IT cortex. The model certainly does not account for all possible aspects of visual perception or illusions (see also extensions, predictions, and future directions below). However, the success of the model in explaining experimental data across multiple scales and making quantitative predictions strongly suggests that the theory provides an important framework for the investigation of the feedforward path in visual cortex and the processes involved in immediate recognition.

An important component of a theory is that it should be falsifiable. In that spirit, we list some key experiments and findings here that could refute the present framework. First, a strong dissociation between experimental observations and model predictions would suggest that revisions need to be made to the model (e.g., psychophysical or physiological observations that cannot be explained or contradict predictions made by the model). Second, as stated in the introduction, the present framework relies entirely on a feedforward architecture from V1 to IT and PFC. Any evidence that feedback plays a key role *during the early stages* of immediate recognition should be considered as hard evidence suggesting that important revisions would need to be made in the main architecture of the model (Fig. 1).

A wish-list of experiments

Here we discuss some predictions from the theory and an accompanying "wish list" of experiments that could be done to test, refute, or validate those predictions. We try to focus on what we naively think are feasible experiments.

1. The distinction between simple and complex cells has been made only in primary visual cortex. Our theory and parsimony considerations suggest that a similar circuit is repeated throughout visual cortex. Therefore, *unbiased* recordings from neurons in higher visual areas may reveal the existence of two classes of neurons which could be distinguished by their degree of invariance to image transformations.
2. As the examples discussed in this manuscript illustrate, our theory can make quantitative predictions about the limits of immediate recognition at the behavioral level (section on "Performance on natural images") and also at the neuronal level (section on "Comparison with physiological observations"). The biggest challenges to recognition include conditions in which the objects are small relative to the whole image and the presence of multiple objects, background, or clutter. It would be interesting to compare these predictions to behavioral and physiological measurements. This could be achieved by adding extra conditions in the psychophysical experiment of section on "Performance on natural images" and by extending the read-out experiments from section "Comparison with physiological observations" to natural images and more complex recognition scenarios.
3. The theory suggests that immediate recognition may rely on a large dictionary of shape-components (i.e., common image-features) with different levels of complexity and invariance. This fits well with the concept of "unbound features" (Treisman and Gelade, 1980; Wolfe and Bennett, 1997) postulated by cognitive theories of pre-attentive vision. Importantly, the theory does not rely on any figure-ground segregation. This suggests that, at least for immediate recognition, recognition can work without an intermediate segmentation step. Furthermore, it also suggests that it is not necessary to define *objects* as fundamental units in visual recognition.
4. There is no specific computational role for a functional topography of units in the model. Thus, the strong degree of topography present throughout cortex, may arise from developmental reasons and physical constraints (a given axon may be more likely to target two adjacent neurons than two neurons that are far away; also, there may be a strong pressure to minimize wiring) as opposed to having a specific role in object recognition or the computations made in cortex.

5. The response of a given simple unit in the model can be described by Eq. (2). Thus, there are multiple *different* inputs that could activate a particular unit. This may explain the somewhat puzzling observations of why physiologists often find neurons that seem to respond to apparently dissimilar objects. Following this reasoning, it should be possible to generate an iso-response stimulus set, i.e., a series of stimuli that should elicit similar responses in a given unit even when the stimuli apparently look different or the shape of the iso-response stimulus set appear non-intuitive.
6. It is tempting to anthropomorphize the responses of units and neurons. This has been carried as far as to speak of a neuron's "preferences." The current theory suggests that an input that gives rise to a high response from a neuron is at the same time simpler and more complex than this anthropomorphized account. It is simpler because it can be rigorously approximated by specific simple equations that control its output. It is more complex because these weight vectors and equations are not easily mapped to words such as "face neuron," "curvature," etc., and taken with the previous point, that visually dissimilar stimuli can give rise to similar responses, the attribution of a descriptive word may not be unique.
7. There are many tasks that may not require back-projections. The performance of the model may provide a reliable signature of whether a task can be accomplished during immediate recognition in the absence of feedback (e.g., the model performs well for immediate recognition of single objects on uncluttered backgrounds, but fails for attention-demanding tasks Li et al., 2002). As stated above, one of the main assumptions of the current model is the feedforward architecture. This suggests that the model may not perform well in situations that require multiple fixations, eye movements, and feedback mechanisms. Recent psychophysical work suggests that performance on dual tasks can provide a diagnostic tool for characterizing tasks that do or do not involve attention (Li et al., 2002). Can the model perform these dual tasks when psychophysics suggests that attention is or is not required? Are back-projections and feedback required?

In addition to the predictions listed above, we recently discussed other experiments and predictions that are based on a more detailed discussion of the biophysical circuits implementing the main operations in the model (see Serre et al., 2005).

Future directions

We end this article by reflecting on several of the open questions, unexplained phenomena, and missing components of the theory. Before we begin, we should note that visual recognition encompasses much more than what has been attempted and achieved with the current theory. A simple example may illustrate this point. In the animal categorization task discussed in the previous sections, humans make mistakes upon being pressed to respond promptly. Given 10 s and no mask, performance would be basically 100%. As stated several times, the goal here is to provide a framework to quantitatively think about the initial steps in vision, but it is clear that much remains to be understood beyond immediate recognition.

Open questions

How strict is the hierarchy and how precisely does it map into cells of different visual areas? For instance, are cells corresponding to S_2 units in V2 and C_2 units in V4 or are some cells corresponding to S_2 units already in V1? The theory is rather open about these possibilities: the mapping of Fig. 1 is just an educated guess. However, because of the increasing arborization of cells and the number of boutons from V1 to PFC (Elston, 2003), the number of subunits to the cells should increase and thus their potential size and complexity. In addition, C units should show more invariance from the bottom to the top of the hierarchy.

What is the nature of the cortical and subcortical connections (both feedforward and feedback) to and *from the main areas of the ventral visual stream that are involved in the model?* A more thorough characterization at the anatomical level of the circuits in visual cortex would lead to a more realistic architecture of the model by better constraining some of the parameters such as the size of the dictionary of shape-components or the number of inputs to units in different layers. This would also help refine and extend the existing literature on the organization of visual cortex (Felleman and van Essen, 1991). With the recent development of higher resolution tracers (e.g., PHA-L, biocytin, DBA), visualization has greatly improved and it is now possible to go beyond a general layout of interconnected structures and start addressing the finer organization of connections.

What are the precise biophysical mechanisms for the learning rule described in section "A quantitative framework for the ventral stream" and how can invariances be learned within the same framework? Possible synaptic mechanisms for learning should be described in biophysical detail. As suggested earlier, synaptic learning rules should allow for three types of learning: (1) the TUNING of the units at the *S* level by detecting correlations among subunits at the same time, (2) the invariance to position and scale at the *C* level by detecting correlations among subunits across time, and (3) the training of task-specific circuits (probably from IT to PFC) in a supervised fashion.

Is learning in areas below IT purely unsupervised and developmental-like as assumed in the theory? Or is there task- and/or object-specific learning in adults occurring below IT in V4, V2, or even V1?

Have we reached the limit of what feedforward architectures can achieve in terms of performance? In other words, is the somewhat better performance of humans on the animal vs. non-animal categorization task (see section on "Comparison between the model and human observers") over the model for *SOAs* longer than 80 ms due to feedback effects mediated by back-projections or can the model be improved to attain human performance in the absence of a mask? There could be several directions to follow in order to try to improve the model performance. One possibility would involve experimenting with the size of the dictionary of shape-components (that could be further reduced with feature selection techniques for instance). Another possibility would involve adding intermediate layers to the existing ones.

Are feedback loops always desirable? Is the performance on a specific task guaranteed to always increase when subjects are given more time? Or are there tasks for which blocking the effect of back-projections with rapid masked visual presentation increases the level of performance compared to longer presentation times?

Future extensions

Learning the tuning of the S_1 units: In the present implementation of the model the tuning of the simple cells in V1 is hardwired. It is likely that it could be determined through the same passive learning mechanisms postulated for the S_2, S_{2b}, and S_3 units (in V4 and PIT respectively), possibly with a slower time scale and constrained to LGN center-surround subunits. We would expect the automatic learning from natural images mostly of oriented receptive fields but also of more complex ones, including end-stopping units [as reported for instance in DeAngelis et al. (1992) in layer 6 of V1].

Dynamics of neuronal responses: The current implementation is completely static, for a given static image the model produces a single response in each unit. This clearly does not account for the intricate dynamics present in the brain and also precludes us from asking several questions about the encoding of visual information, learning, the relative timing across areas, etc. Perhaps the easiest way to solve this is by using simple single neuron models (such as an integrate-and-fire neuron) for the units in the model. This question is clearly related to the biophysics of the circuitry, i.e., what type of biological architectures and mechanisms can give rise to the global operations used by the model. A dynamical model would allow us to more realistically compare to experimental data. For example, the experiments described in section "Performance on natural images" compare the results in a categorization task between the model and human subjects. In the human psychophysics, the stimuli were masked briefly after stimulus presentation. A dynamical model would allow us to investigate the role and mechanisms responsible for masking. A dynamical model may also allow investigation of time-dependent phenomena as well as learning based on correlations across time.

Extensions of the model to other visual inputs: There are many aspects of vision that are not currently implemented in the model. These include color, stereo, motion, and time-varying stimuli. Initial work has been done to extend the model to the visual recognition of action and motions (Giese and Poggio, 2003; Sigala et al., 2005). It is likely that the same units supporting recognition of static images (the S_4, view-tuned units in the model) show time sequence selectivity.

Color mechanisms from V1 to IT should be included. The present implementation only deals with gray level images (it has been shown that the addition of color information in rapid categorization tasks only leads to a mild increase in performance Delorme et al., 2000). More complex phenomena involving color such as color constancy and the influence of the background and integration in color perception should ultimately be explained.

Stereo mechanisms from V1 to IT should also be included. Stereo and especially motion play an important role in the learning of invariances such as position and size invariance via a correlation-based rule such as the trace rule (Földiák, 1991).

Extensions of the anatomy of the model: Even staying within the feedforward skeleton outlined here, there are many connections that are known to exist in the brain that are not accounted for in the current model. The goal of the model is to extract the basic principles used in recognition and not to copy, neuron by neuron, the entire brain. However, certain connectivity patterns may have important computational consequences. For example, there are horizontal connections in the cortex that may be important in modulating and integrating information across areas beyond the receptive field.

Beyond a feedforward model: It has been known for many decades now that there are abundant back-projections in the brain. In the visual system, every area projects back to its input area (with the exception of the lateral geniculate nucleus in the thalamus that does not project back to the retina). Some of these connections (e.g., from V2 to V1), may play a role even during immediate recognition. However, a central assumption of the current model is that long-range backprojections (e.g., from area IT to V1) do not play a role during the first 100–150 ms of vision. Given enough time, humans make eye movements to scan an image and performance in many object recognition tasks can increase significantly over that obtained during fast presentation.

Visual illusions: A variety of visual illusions show striking effects that are often counterintuitive and require an explanation in terms of the neuronal circuits. While in some cases specific models have been proposed to explain one phenomenon or another, it would be interesting to explore how well the model (and thus feedforward vision) can account for those observations. A few simple examples include illusory contours (such as the Kanizsa triangle), long-range integration effects (such as the Cornsweet illusion), etc. More generally, it is likely that early Gestalt-like mechanisms — for detecting collinearity, symmetry, parallelism, etc. — exist in V1 or V2 or V4. They are not present in this version of the model. It is an open and interesting question how they could be added to it in a plausible way.

Acknowledgments

We would like to acknowledge Chou Hung and James DiCarlo for their contributions in the physiology experiments described in section "Comparison with physiological observations." We would like to acknowledge Aude Oliva for her contribution to the psychophysics experiments described in section "Performance on natural images." We thank Timothee Masquellier for useful comments on this manuscript. This research was sponsored by grants from: Office of Naval Research (DARPA) under Contract No. N00014-00-1-0907, McGovern fellowship (GK), National Science Foundation (ITR) under Contract No. IIS-0085836, National Science Foundation (KDI) under Contract No. DMS-9872936, and National

Science Foundation under Contract No. IIS-9800032.

References

Abbott, L.F., Rolls, E.T. and Tovee, M.T. (1996) Representational capacity of face coding in monkeys. Cereb. Cortex, 6: 498–505.

Amit, Y. and Mascaro, M. (2003) An integrated network for invariant visual detection and recognition. Vision Res., 43(19): 2073–2088.

Bacon-Mace, N., Mace, M.J., Fabre-Thorpe, M. and Thorpe, S.J. (2005) The time course of visual processing: backward masking and natural scene categorisation. Vision Res., 45: 1459–1469.

Barlow, H.B. (1961) Possible principles underlying the transformation of sensory messages. In: Rosenblith W.A. (Ed.), Sensory Communication. MIT Press, Cambridge, MA, pp. 217–234.

Biederman, I. (1987) Recognition-by-components: a theory of human image understanding. Psychol. Rev., 94: 115–147.

Booth, M.C. and Rolls, E.T. (1998) View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cereb. Cortex, 8: 510–523.

AU :4 Breitmeyer, B. and Ogmen, H. (2006) Visual Masking: Time Slices through Conscious and Unconscious Vision. Oxford University Press.

DeAngelis, G.C., Robson, J.G., Ohzawa, I. and Freeman, R.D. (1992) Organization of suppression in receptive fields of neurons in cat visual cortex. J. Neurophysiol., 68(1): 144–163.

Delorme, A., Richard, G. and Fabre-Thorpe, M. (2000) Ultra-rapid categorisation of natural images does not rely on colour: a study in monkeys and humans. Vision Res., 40: 2187–2200.

Desimone, R., Albright, T.D., Gross, C.G. and Bruce, C. (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. J. Neurosci., 4(8): 2051–2062.

Elston, G.N. (2003) Comparative studies of pyramidal neurons in visual cortex of monkeys. In: Kaas J.H. and Collins C. (Eds.), The Primate Visual System. CRC Press, Boca Raton, FL, pp. 365–385.

Enns, J.T. and Di Lollo, V. (2000) What's new in masking? Trends Cogn. Sci., 4(9): 345–351.

AU :5 Fei-Fei, L., Fergus, R. and Perona, P. (2004) Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: Proc. IEEE CVPR, Workshop on generative-model based vision.

Felleman, D.J. and van Essen, D.C. (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex, 1: 1–47.

Földiák, P. (1991) Learning invariance from transformation sequences. Neural Comput., 3: 194–200.

Freedman, D.J., Riesenhuber, M., Poggio, T. and Miller, E.K. (2002) Visual categorization and the primate prefrontal cortex: neurophysiology and behavior. J. Neurophysiol., 88: 930–942.

Freiwald, W.A., Tsao, D.Y., Tootell, R.B.H. and Livingstone, M.S. (2005) Complex and dynamic receptive field structure in macaque cortical area V4d. J. Vis., 4(8): 184a.

Fukushima, K. (1980) Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern., 36: 193–202.

Gallant, J.L., Connor, C.E., Rakshit, S., Lewis, J.W. and Van Essen, D.C. (1996) Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. J. Neurophysiol., 76: 2718–2739.

Gawne, T.J. and Martin, J.M. (2002) Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. J. Neurophysiol., 88: 1128–1135.

Giese, M. and Poggio, T. (2003) Neural mechanisms for the recognition of biological movements and action. Nat. Rev. Neurosci., 4: 179–192.

Gross, C.G. (1998) Brain Vision and Memory: Tales in the History of Neuroscience. MIT Press.

Gross, C.G., Rocha-Miranda, C.E. and Bender, D.B. (1972) Visual properties of neurons in inferotemporal cortex of the macaque. J. Neurophysiol., 35: 96–111.

Hegdé, J. and van Essen, D.C. (2006) A comparative study of shape representation in macaque visual areas V2 and V4. Cereb. Cortex. AU :6

Hietanen, J.K., Perrett, D.I., Oram, M.W., Benson, P.J. and Dittrich, W.H. (1992) The effects of lighting conditions on responses of cells selective for face views in the macaque temporal cortex. Exp. Brain Res., 89: 157–171.

Hubel, D.H. and Wiesel, T.N. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol., 160: 106–154.

Hung, C., Kreiman, G., Poggio, T. and DiCarlo, J. (2005) Fast read-out of object identity from macaque inferior temporal cortex. Science, 310: 863–866.

Keysers, C., Xiao, D.K., Földiák, P. and Perrett, D.I. (2001) The speed of sight. J. Cogn. Neurosci., 13: 90–101.

Kobatake, E. and Tanaka, K. (1994) Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J. Neurophysiol., 71: 856–867.

Kreiman, G., Hung, C., Poggio, T. and DiCarlo, J. (2006) Object selectivity of local field potentials and spikes in the inferior temporal cortex of macaque monkeys. Neuron, 49: 433–445.

Lamme, V.A.F. and Roelfsema, P.R. (2000) The disctinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci., 23: 571–579.

Lampl, I., Ferster, D., Poggio, T. and Riesenhuber, M. (2004) Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. J. Neurophysiol., 92: 2704–2713.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-based learning applied to document recognition. Proc. IEEE, 86(11): 2278–2324.

Li, F.F., VanRullen, R., Koch, C. and Perona, P. (2002) Rapid natural scene categorization in the near absence of attention. Proc. Natl. Acad. Sci. U.S.A., 99: 9596–9601.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T. and Oeltermann, A. (2001) Neurophysiological investigation of the basis of the fMRI signal. Nature, 412: 150–157.

Logothetis, N.K., Pauls, J. and Poggio, T. (1995) Shape representation in the inferior temporal cortex of monkeys. Curr. Biol., 5: 552–563.

Logothetis, N.K. and Sheinberg, D.L. (1996) Visual object recognition. Ann. Rev. Neurosci., 19: 577–621.

Mahon, L.E. and DeValois, R.L. (2001) Cartesian and non-Cartesian responses in LGN, V1, and V2 cells. Vis. Neurosci., 18: 973–981.

Mel, B.W. (1997) SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. Neural Comput., 9: 777–804.

Miller, E.K. (2000) The prefrontal cortex and cognitive control. Nat. Rev. Neurosci., 1: 59–65.

Mitzdorf, U. (1985) Current source-density method and application in cat cerebral cortex: investigation of evoked potentials and EEG phenomena. Physiol. Rev., 65: 37–99.

Mutch, J. and Lowe, D. (2006) Multiclass object recognition with sparse, localized features. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.

Nakamura, H., Gattass, R., Desimone, R. and Ungerleider, L.G. (1993) The modular organization of projections from areas V1 and V2 to areas V4 and TEO in macaques. J. Neurosci., 13(9): 3681–3691.

Olshausen, B.A., Anderson, C.H. and Van Essen, D.C. (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neurosci., 13(11): 4700–4719.

Pasupathy, A. and Connor, C.E. (2001) Shape representation in area V4: position-specific tuning for boundary conformation. J. Neurophysiol., 86(5): 2505–2519.

Perrett, D.I., Hietanen, J.K., Oram, M.W. and Benson, P.J. (1992) Organization and functions of cells responsive to faces in the temporal cortex. Philos. Trans. R. Soc. B, 335: 23–30.

Perrett, D.I. and Oram, M. (1993) Neurophysiology of shape processing. Image Vis. Comput., 11: 317–333.

Poggio, T. and Bizzi, E. (2004) Generalization in vision and motor control. Nature, 431: 768–774.

Poggio, T. and Edelman, S. (1990) A network that learns to recognize 3D objects. Nature, 343: 263–266.

Potter, M.C. (1975) Meaning in visual search. Science, 187: 565–566.

Reynolds, J.H., Chelazzi, L. and Desimone, R. (1999) Competitive mechanisms subserve attention in macaque areas V2 and V4. J. Neurosci., 19: 1736–1753.

Riesenhuber, M. and Poggio, T. (1999) Hierarchical models of object recognition in cortex. Nat. Neurosci., 2: 1019–1025.

Rolls, E.T., Tovee, M.J. and Panzeri, S. (1999) The neurophysiology of backward visual masking: information analysis. J. Comp. Neurol., 11: 300–311.

Rousselet, G.A., Mace, M.J. and Fabre-Thorpe, M. (2003) Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. J. Vis., 3: 440–455.

Sato, T. (1989) Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake macaques. Exp. Brain Res., 74(2): 263–271.

Schwartz, E.L., Desimone, R., Albright, T.D. and Gross, C.G. (1983) Shape recognition and inferior temporal neurons. Proc. Natl. Acad. Sci. U.S.A., 80(18): 5776–5778.

Serre, T. (2006) Learning a dictionary of shape-components in visual cortex: comparison with neurons, humans and machines. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, April 2006.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G. and Poggio, T. (2005) A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. AI Memo 2005-036/CBCL Memo 259, MIT, Cambridge, MA.

Serre, T., Oliva, A. and Poggio, T. (2007a) A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. (in press).

Serre, T. and Riesenhuber, M. (2004) Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex. AI Memo 2004-017/CBCL Memo 239, MIT, Cambridge, MA.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M. and Poggio, T. (2007b) Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Anal. Machine Intell., 29(3): 411–426.

Sigala, R., Serre, T., Poggio, T. and Giese, M. (2005) Learning features of intermediate complexity for the recognition of biological motion. In: Proc. Int. Conf. Artif. Neural Netw.

Simons, D.J. and Rensink, R.A. (2005) Change blindness: past, present and future. Trends Cogn. Sci., 9(1): 16–20.

Tanaka, K. (1996) Inferotemporal cortex and object vision. Ann. Rev. Neurosci., 19: 109–139.

Thorpe, S.J. (2002) Ultra-rapid scene categorisation with a wave of spikes. In: Proc. Biologically Motivated Comput. Vis.

Thorpe, S.J. and Fabre-Thorpe, M. (2001) Seeking categories in the brain. Science, 291: 260–263.

Thorpe, S.J., Fize, D. and Marlot, C. (1996) Speed of processing in the human visual system. Nature, 381: 520–522.

Torralba, A. and Oliva, A. (2003) Statistics of natural image categories. Netw Comput. Neural Syst., 14: 391–412.

Tovee, M.J., Rolls, E.T., Treves, A. and Bellis, R.P. (1993) Information encoding and the response of single neurons in the primate temporal visual cortex. J. Neurophysiol.

Treisman, A.M. and Gelade, G. (1980) A feature-integration theory of attention. Cogn. Psychol., 12: 97–136.

Ullman, S., Vidal-Naquet, M. and Sali, E. (2002) Visual features of intermdediate complexity and their use in classification. Nat. Neurosci., 5(7): 682–687.

Ungerleider, L.G. and Haxby, J.V. (1994) ‘‘What’’ and ‘‘where’’ in the human brain. Curr. Opin. Neurobiol., 4: 157–165.

Valiant, L.G. (2005) Memorization and association on a realistic neural model. Neural Comput., 17: 527–555.

VanRullen, R. and Koch, C. (2003) Visual selective behavior can be triggered by a feed-forward process. J. Comp. Neurol., 15: 209–217.

Victor, J.D., Mechler, F., Repucci, M.A., Purpura, K.P. and Sharpee, T. (2006) Responses of V1 neurons to two-dimensional hermite functions. J. Neurophysiol., 95: 379–400.

Wallis, G. and Rolls, E.T. (1997) A model of invariant object recognition in the visual system. Prog. Neurobiol., 51: 167–194.

Wersing, H. and Koerner, E. (2003) Learning optimized features for hierarchical models of invariant recognition. Neural Comput., 15(7): 1559–1588.

Wolfe, J.M. and Bennett, S.C. (1997) Preattentive object files: shapeless bundles of basic features. Vision Res., 37: 25–44.